

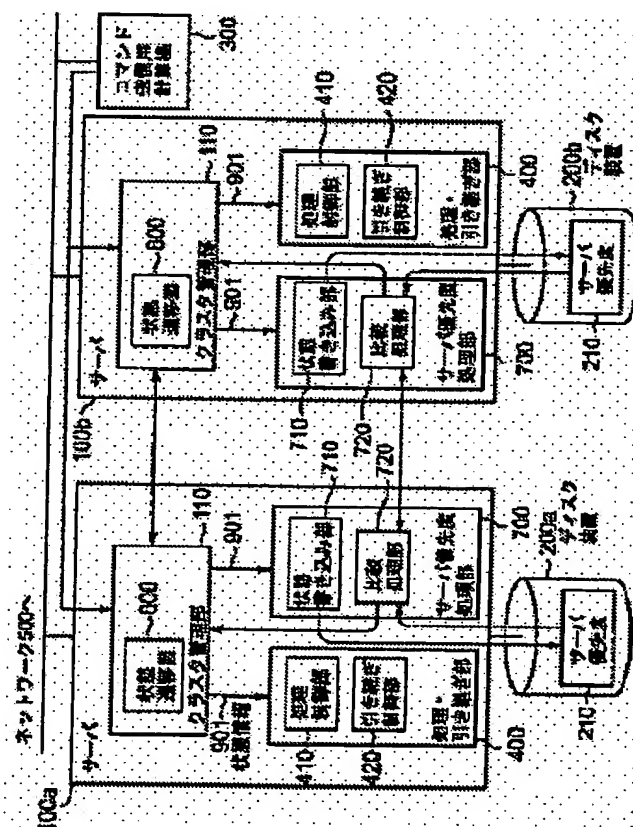
RECORDING MEDIUM WITH RECORDED PROGRAM DECIDING SERVER COMPUTER HAVING PERFORMED PROCESSING FINALLY AND HIGH-AVAILABILITY COMPUTER SYSTEM

Patent number: JP2001175626
 Publication date: 2001-06-29
 Inventor: YAMAMOTO KOJI; ENDO KOTARO
 Applicant: TOKYO SHIBAURA ELECTRIC CO
 Classification:
 - international: G06F15/177; G06F11/20
 - european:
 Application number: JP19990364571 19991222
 Priority number(s): JP19990364571 19991222

Report a data error here

Abstract of JP2001175626

PROBLEM TO BE SOLVED: To actualize automatic operation by eliminating the limit of the frequency of hand-over of a processing and to improve decision precision by deciding a server having performed a process finally by using only local information. **SOLUTION:** When a fault occurs in at least one server 100a or 100b and when the server recovers from the fault, the cluster management part 110 of the server shifts server state according to a state transition chart 800. A state write part 710 records server priority 210 determined by the server state that state change information 901 indicates on disk drivers 200a and 200b. A comparing processing part 720 decides whether or not its server has higher priority according to at least the priority 210 of its server if a fault occurs to the servers 100a and 100b and at least its server recovers from the fault thereafter. The cluster management part 110 makes a corresponding state shift according to the decision result and state transition chart 800.



Data supplied from the esp@cenet database - Worldwide

【特許請求の範囲】

【請求項1】 2台のサーバ計算機的一方が処理を行い、処理を行っているサーバ計算機で障害が発生した場合に、もう一方のサーバ計算機が処理を引き継ぐことが可能な高可用性計算機システムに適用され、前記各サーバ計算機に実行させるための最後に処理を行っていたサーバ計算機を判定するプログラムを記録した計算機読み取り可能な記録媒体であって、

前記各サーバ計算機に、

前記2台のサーバ計算機の少なくとも一方で障害が発生した場合と障害から復帰した場合とに、所定の状態遷移図に従って当該サーバ計算機の状態遷移を行う状態遷移ステップと、

前記状態遷移に応じて、少なくとも当該遷移後の自サーバ計算機の状態が決まるサーバ優先度を自サーバ計算機がローカルに持つ記憶装置に記録するサーバ優先度記録ステップと、

前記2台のサーバ計算機の両方に障害が発生し、その後少なくとも自サーバ計算機が障害から復帰した際に、少なくとも当該自サーバ計算機の前記記憶装置に記録されたサーバ優先度に基づいて当該自サーバ計算機の方が優先度が高いか否かを判定する優先度判定ステップと、前記優先度判定ステップでの優先度判定結果に基づいて、自サーバ計算機が最後に処理を行っていたサーバ計算機であるか否かを判定することで、自サーバ計算機が処理を継続するか否かを決定し、その決定結果により前記状態遷移ステップでの状態遷移を行わせる最終処理計算機判定ステップとを実行させるための最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体。

【請求項2】 前記優先度判定ステップは、前記2台のサーバ計算機の両方に障害が発生し、その後両サーバ計算機が共に障害から復帰した際に、当該両サーバ計算機の前記記憶装置に記録されたサーバ優先度を比較して、自サーバ計算機の方が優先度が高いか否かを判定する第1の優先度判定ステップを含むことを特徴とする請求項1記載の最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体。

【請求項3】 前記優先度判定ステップは、前記2台のサーバ計算機の両方に障害が発生し、その後自サーバ計算機のみが障害から復帰した際に、自サーバ計算機の前記記憶装置に記録されたサーバ優先度が最高優先度であるか否かを判定する最高優先度判定ステップと、前記最高優先度判定ステップで最高優先度であると判定された場合にのみ自サーバ計算機の方が優先度が高いと判定する第2の優先度判定ステップを更に含むことを特徴とする請求項2記載の最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体。

【請求項4】 前記状態遷移ステップは、前記各サーバ計算機の状態を、処理を行い、且つ行っている処理を引

き継ぐ相手が存在する「マスタ」と、処理は行いが、行っている処理を引き継ぐ相手が存在しない「シングルマスタ」と、処理を行っていないが、引き継ぎのための情報を受け取っている「スレーブ」と、処理を行っておらず、且つ引き継ぎのための情報を受け取っていない「停止」の4つの状態に分類して作成された前記状態遷移図に従って前記2台のサーバ計算機の状態遷移を行うことを特徴とする請求項3記載の最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体。

【請求項5】 前記状態遷移ステップは、前記2台のサーバ計算機の両方に障害が発生し、その後両サーバ計算機が共に障害から復帰した際に、前記最終処理計算機判定ステップでの判定結果と前記状態遷移図とに基づいて、前記2台のサーバ計算機的一方が前記停止状態から前記シングルマスタ状態となり、他方が前記停止状態から前記スレーブ状態となる状態遷移を行う第1の状態遷移ステップと、その後前記他方がスレーブ状態のまま、前記一方がマスタ状態となる状態遷移を行う第2の状態遷移ステップとを含むことを特徴とする請求項4記載の最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体。

【請求項6】 前記状態遷移ステップは、前記2台のサーバ計算機の両方に障害が発生し、その後自サーバ計算機のみが障害から復帰した際に、前記最終処理計算機判定ステップでの判定結果と前記状態遷移図とに基づいて、自サーバ計算機が前記停止状態から前記シングルマスタ状態となる状態遷移を行うか、或いは現在の状態を維持する第3の状態遷移ステップと、その後相手サーバ計算機が障害から復帰した際に、当該相手サーバ計算機が前記停止状態から前記スレーブ状態となる状態遷移を行うか、或いは前記2台のサーバ計算機的一方が前記シングルマスタ状態となり、他方が前記スレーブ状態となる状態遷移を行う第4の状態遷移ステップとを含むことを特徴とする請求項4記載の最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体。

【請求項7】 前記状態遷移ステップは、前記2台のサーバ計算機の両方に障害が発生し、その後自サーバ計算機のみが障害から復帰した際に、当該自サーバ計算機の前記記憶装置に記録されたサーバ優先度が最高優先度でないために、前記最高優先度判定ステップでの判定ができず、当該自サーバ計算機が処理を継続する計算機であると判定できない状態で、当該自サーバ計算機に対して強制的に処理を継続させる強制開始命令が外部から与えられた場合に、当該自サーバ計算機が処理を継続するサーバ計算機として当該自サーバ計算機の状態遷移を行わせる強制開始ステップを含むことを特徴とする請求項3記載の最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体。

【請求項8】 2台のサーバ計算機的一方が処理を行い、処理を行っているサーバ計算機で障害が発生した場

合に、もう一方のサーバ計算機が処理を引き継ぐことが可能な高可用性計算機システムにおいて、

前記各サーバ計算機は、

前記2台のサーバ計算機の少なくとも一方で障害が発生した場合と障害から復帰した場合とに、所定の状態遷移図に従って当該サーバ計算機の状態遷移を行う状態遷移手段と、

前記状態遷移手段により状態遷移が行われる度に、少なくとも当該遷移後の自サーバ計算機の状態が決まるサーバ優先度を自サーバ計算機がローカルに持つ記憶装置に記録する状態書き込み手段と、

前記2台のサーバ計算機の両方に障害が発生し、その後少なくとも自サーバ計算機が障害から復帰した際に、少なくとも当該自サーバ計算機の前記記憶装置に記録されたサーバ優先度に基づいて当該自サーバ計算機の方が優先度が高いか否かを判定する優先度判定手段と、

前記優先度判定手段での優先度判定結果に基づいて、自サーバ計算機が最後に処理を行っていたサーバ計算機であるか否かを判定することで、自サーバ計算機が処理を継続するか否かを決定し、その決定結果により前記状態遷移手段による状態遷移を行わせる最終処理計算機判定手段とを具備することを特徴とする計算機システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、2台のサーバ計算機の一方が処理を行い、処理を行っているサーバ計算機で障害が発生しても、もう一方のサーバ計算機が処理を引き継ぐことができる高可用性計算機システムに係り、特に両方のサーバ計算機に障害が発生し、その後少なくとも一方が障害から復帰した場合に、その復帰した計算機が処理を継続するか否かを判定するのに必要な、最後に処理を行っていたサーバ計算機を判定するプログラムを記録した記録媒体、及び高可用性計算機システムに関する。

【0002】

【従来の技術】従来から、複数のサーバ計算機（以下、サーバと略称する）、例えば2台のサーバをネットワーク等で結合し、1つのサーバで障害が発生しても、障害で停止した処理（サービス）を別のサーバが引き継ぐことにより、システム全体として可用性を維持できるようにしたクラスタ型の耐障害コンピュータシステム、つまり高可用性計算機システムが種々開発されている。

【0003】この種の計算機システムのうち、共有ディスク装置等の共有記憶装置を備えたシステムでは、一方のサーバで処理を実行しているときに、当該サーバから処理を引き継ぐのに必要な情報を当該共有記憶装置に記録しておくのが一般的である。

【0004】このような構成の計算機システムでは、例えば2台のサーバに障害が発生して、その後両サーバが共に障害から復帰した場合、或いは、いずれか一方のサ

サーバだけが障害から復帰した場合のいずれの場合にも、共有記憶装置に記録された情報を利用することで、どの障害復帰サーバであっても容易に引き継ぎを行うことができる。

【0005】ところが、高可用性計算機システムの中には、共有記憶装置を持たないシステムもある。このような計算機システムでは、処理の引き継ぎを可能とするために、一方のサーバで処理を実行しているときに、当該サーバから他方のサーバに対し、処理の続きをするのに必要な情報を送るのが一般的である。

【0006】このようにすると、処理を実行しているサーバに障害が発生して処理を続行できなくなった場合に、他方のサーバでそれまでに送られてきた情報を使って処理の続きを行うこと、つまり処理を引き継ぐことが可能となる。

【0007】しかし、両方のサーバに障害が発生した場合に上記の処理の引き継ぎを行うのは容易ではない。その理由は、例えば両サーバが共に障害復帰（復旧）した場合であれば、いずれのサーバが処理を行う（引き継ぐ）かを決定しなければならないからである。また、いずれか一方のサーバだけが障害から復帰した場合であれば、その復帰したサーバで処理を行う（引き継ぐ）かを判定しなければならないからである。この従来技術について、以下に詳述する。

【0008】2台のサーバに障害が発生して、両サーバが共に障害復帰した場合、それまでに処理を行っていなかったサーバ（スレーブ）が処理の続きを行うためには、処理を行っていたサーバ（マスタ）から引き継ぎのための情報をもらっておく必要がある。

【0009】ところが、最後に処理を行っていたサーバ（マスタ）に障害が発生する直前に、このサーバからもう一方のサーバ（スレーブ）に対して、引き継ぎのための情報を送っていなかった場合は、当該もう一方のサーバ（スレーブ）では処理の続きを行うことができない。そのため、両サーバが共に障害復帰した際には、最後に処理を行っていた方のサーバを、処理を引き継ぐサーバとして決定（選択）する必要がある。

【0010】次に、最後に処理を行っていたサーバ（マスタ）に障害が発生する前に、このサーバからもう一方のサーバ（スレーブ）に対して、引き継ぎのための情報を送っていた場合は、両サーバが共に障害復帰した際に、どちらのサーバを選択しても処理の続きをすることが可能である。しかし、障害発生直前の処理については、最後に処理を行っていたサーバから引き継ぎに必要な情報をもう一台のサーバに送る前に障害が発生している可能性がある。そのため、こちらの場合も、最後に処理を行っていた方のサーバを選択する必要がある。

【0011】また、いずれか一方のサーバだけが障害から復帰した場合にも、そのサーバに処理をさせる条件は、両サーバが共に障害復帰した場合と同様の理由によ

り、そのサーバが2台のうち最後まで処理を行っていたサーバであることである。

【0012】そのため従来は、最後に処理を行っていたサーバが判定可能なように、次に述べる2つの方法のいずれかを適用していた。

【0013】(1) 引き継ぎを1度に限定する方法
 予め、どちらか1つのサーバをプライマリサーバ、もう1つのサーバをセカンダリサーバと決めておき、初めは必ずプライマリサーバがマスタ、セカンダリサーバがスレーブとして動作を開始する。ここで、マスタは処理（クライアントから要求された処理）を行い、引き継ぎのための情報をスレーブに送る。スレーブはマスタから送られる引き継ぎのための情報を受け取って自身の有するディスク装置等の外部記憶装置（ローカルな外部記憶装置）に保存する。

【0014】そして、1度、プライマリサーバに障害が発生して引き継ぎを行ったなら、その後はプライマリサーバが復帰しても利用しない。つまり、引き継ぎを1度に限定する。この場合、セカンダリサーバが、自身の外部記憶装置に自身が処理を行ったかどうかを記録しておくことにすれば、容易にどちらが最近まで処理を行っていたかを判定できる。

【0015】しかし、この従来技術は引き継ぎが1度しかできないため、片方または両方のサーバに障害発生、或いは障害復帰がいつ起こっても、可能な限り処理を継続するという、いわゆる自動運転が実現できない。

【0016】(2) 時刻情報を使用する方法
 予め、2台のサーバの時計（時刻）を合わせておく。サーバが処理を開始するときに、現在時刻を自身の持つ外部記憶装置に記録する。このようにすると、両サーバが共に障害復帰した際に、外部記憶装置に記録した時刻の情報をネットワークを通じて互いに授受し、より新しい時刻情報を持つサーバを、最後に処理を行っていたサーバとして判定する。

【0017】この時刻情報を使う方法は、時間というグローバルなものを暗黙に仮定しており、各サーバの使う時計が常に同期しているということを仮定している。しかし、実際の時計は必ずしも常に同期しているとは限らず、この方法は完全性に問題がある。また、一方のサーバだけが障害復帰した場合には、もう一方のサーバとの間で時刻情報の授受が行えないため、自身が最後に処理を行っていたサーバであるか否か判定できない。

【0018】

【発明が解決しようとする課題】 上記したように、2台のサーバの一方が処理を行い、処理を行っているサーバで障害が発生しても、もう一方のサーバが処理を引き継ぐことが可能な高可用性計算機システムであって、共有記憶装置を持たない従来の高可用性計算機システムでは、2台のサーバに障害が発生し、その後少なくとも一方が障害から復帰した場合に、処理を継続するサーバ、

即ち最後に処理を行っていたサーバを判定可能とする方法として、「引き継ぎを1度に限定する方法」、或いは「時刻情報を使用する方法」が適用されていた。

【0019】しかし、「引き継ぎを1度に限定する方法」では、引き継ぎが1度しかできないため、自動運転が実現できないという問題があった。また、「時刻情報を使用する方法」では、時間というグローバルなものを暗黙に仮定し、各サーバの使う時計が常に同期しているということを仮定しているが、実際の時計は必ずしも常に同期しているとは限らず、したがって判定精度（完全性）に問題があった。また、一方のサーバだけが障害復帰した場合には、自身が最後に処理を行っていたサーバであるか否か判定できないという問題もあった。

【0020】本発明は上記事情を考慮してなされたものでその目的は、時刻情報といったグローバルな情報を使わず、ローカルな情報だけを使って最後に処理を行っていたサーバを判定することにより、処理引き継ぎの回数の制限をなくして自動運転を実現すると共に、判定精度の向上が図れる、最後に処理を行っていたサーバ（サーバ計算機）を判定するプログラムを記録した記録媒体、及び高可用性計算機システムを提供することにある。

【0021】

【課題を解決するための手段】 本発明の記録媒体は、2台のサーバの一方が処理を行い、処理を行っているサーバで障害が発生した場合に、もう一方のサーバが処理を引き継ぐことが可能な高可用性計算機システムに適用される、最後に処理を行っていたサーバを判定するプログラムであって、上記各サーバに、以下の各ステップ、即ち、上記2台のサーバの少なくとも一方で障害が発生した場合と障害から復帰した場合とに、所定の状態遷移図に従って当該サーバの状態遷移を行う状態遷移ステップと、この状態遷移に応じて、少なくとも当該遷移後の自サーバの状態で決まるサーバ優先度を自サーバがローカルに持つ記憶装置に記録するサーバ優先度記録ステップと、サーバの両方に障害が発生し、その後少なくとも自サーバが障害から復帰した際に、少なくとも当該自サーバの記憶装置に記録されたサーバ優先度に基づいて当該自サーバの方が優先度が高いか否かを判定する優先度判定ステップと、この優先度判定ステップでの優先度判定結果に基づいて、自サーバが最後に処理を行っていたサーバであるか否かを判定することで、自サーバが処理を継続するか否かを決定し、その決定結果により上記状態遷移ステップでの状態遷移を行わせる最終処理計算機判定ステップとを実行させるための、最後に処理を行っていたサーバを判定するプログラムを記録したことを特徴とする。

【0022】このような構成においては、所定の状態遷移図に従う状態の遷移で決まるサーバ優先度という状態変数がローカルな記憶装置に記録され、当該状態遷移とローカルな状態変数を用いて、最後まで処理を行ってい

たサーバが判定されることから、処理引き継ぎの回数の制限をなくして自動運転を実現することが可能となる。

【0023】ここで、上記2台のサーバの両方に障害が発生し、その後両サーバが共に障害から復帰した際の処理のために、上記優先度判定ステップに次のステップ、即ち、上記両サーバが共に障害から復帰した際に、当該両サーバのサーバ優先度を比較して、自サーバの方が優先度が高いか否かを判定する第1の優先度判定ステップを持たせるとよい。

【0024】また、上記2台のサーバの両方に障害が発生し、その後自サーバのみが障害から復帰した際の処理のために、上記優先度判定ステップに次の2つのステップ、即ち、自サーバのサーバ優先度が最高優先度であるか否かを判定する最高優先度判定ステップと、ここで最高優先度であると判定された場合にのみ自サーバの方が優先度が高いと判定する第2の優先度判定ステップとを持たせるとよい。

【0025】また本発明は、上記状態遷移図により示される上記2台のサーバの状態として、処理を行い、且つ行っている処理を引き継ぐ相手が存在する「マスタ」と、処理は行うが、行っている処理を引き継ぐ相手が存在しない「シングルマスタ」と、処理を行っていないが、引き継ぎのための情報を受け取っている「スレーブ」と、処理を行っておらず、且つ引き継ぎのための情報を受け取っていない「停止」の4つの状態の組み合わせを適用したことを特徴とする。

【0026】このように、処理を行うサーバを、行っている処理を引き継ぐ相手が存在する「マスタ」と、行っている処理を引き継ぐ相手が存在しない「シングルマスタ」とに分類して区別することで、両サーバのローカルな状態変数（サーバ優先度）が同一値となるのを防止して、当該ローカルな状態変数を用いて最後に処理を行っていたサーバを判定する場合の判定精度を向上することが可能となる。

【0027】この効果は、上記サーバ優先度記録ステップに次の4つのステップ、即ち、自サーバの状態がシングルマスタ状態に遷移する場合に、自サーバのサーバ優先度を最高優先度を示すように変更する第1のサーバ優先度記録ステップと、自サーバの状態がマスタ状態に遷移する場合に、自サーバのサーバ優先度を2番目の優先度を示すように変更する第2のサーバ優先度記録ステップと、自サーバの状態がスレーブ状態に遷移する場合に、自サーバのサーバ優先度を最低優先度を示すように変更する第3のサーバ優先度記録ステップと、自サーバの状態が停止状態に遷移する場合に、自サーバのサーバ優先度の変更を抑止する（元優先度を再度記録することと等価）サーバ優先度維持ステップとを持たせることにより、一層顕著となる。

【0028】上記4つのサーバ状態を適用した構成では、2台のサーバの両方に障害が発生し、その後両サー

バが共に障害から復帰した際に、上記最終処理計算機判定ステップでの判定結果と状態遷移図とに基づいて、上記2台のサーバの一方が停止状態からシングルマスタ状態となり、他方が停止状態からスレーブ状態となる状態遷移を行い、その後上記他方がスレーブ状態のまま、上記一方がマスタ状態となる状態遷移を行うとよい。

【0029】また、2台のサーバの両方に障害が発生し、その後自サーバのみが障害から復帰した際には、上記最終処理計算機判定ステップでの判定結果と状態遷移図とに基づいて、自サーバが停止状態からシングルマスタ状態となる状態遷移を行うか、或いは現在の状態を維持し、その後相手サーバが障害から復帰した際に、当該相手サーバが停止状態からスレーブ状態となる状態遷移を行うか、或いは一方がシングルマスタ状態となり、他方がスレーブ状態となる状態遷移を行うとよい。

【0030】また本発明は、上記状態遷移ステップに次のステップ、即ち、上記2台のサーバの両方に障害が発生し、その後自サーバのみが障害から復帰した際に、当該自サーバのサーバ優先度が最高優先度でないために、上記最高優先度判定ステップでの判定ができず、当該自サーバが処理を継続する計算機であると判定できない状態で、当該自サーバに対して強制的に処理を継続させる強制開始命令が外部から与えられた場合に、当該自サーバが処理を継続するサーバとして当該自サーバの状態遷移を行わせる強制開始ステップを持たせたことをも特徴とする。

【0031】このような構成においては、2台のサーバに障害が発生し、その後1台だけが障害復帰しても、その1台のサーバ優先度が最高優先度でないために、その1台の方が停止状態にあるもう1台より優先度が高いと判定することが不可能で、したがって、その1台が最後に処理を行っていたサーバであると判定できない場合でも、その1台に対して外部の計算機（コマンド送信用計算機）から強制開始命令を与えることで、強制的に処理を行わせることが可能となる。このように強制的に処理を開始させることができる機能は、処理の続きは無理でも、処理を再開した方が都合がよい場合に便利である。

【0032】ここで、停止状態にあるサーバの優先度が最高優先度のときに、障害復帰したサーバに対して強制的に処理を行わせると、両サーバが共に最高優先度となる。このような状態では、両サーバに共に障害が発生して、その後共に障害復帰した場合に、優先度の高いサーバを特定できず、したがって処理の引き継ぎを自動的に行うことはできないが、再度強制開始を行うことで容易に対処可能となる。

【0033】なお、上記した最後に処理を行っていたサーバを判定するプログラムを記録した記録媒体に係る本発明は、各計算機が当該プログラムで実現される機能手段を備えた高可用性計算機システムに係る発明としても成立し、当該プログラムで適用される手順を持つ方法に

係る発明としても成立する。

【0034】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図1は本発明の一実施形態に係る高可用性計算機システムの構成を示すブロック図である。

【0035】図1のシステムにおいて、ネットワーク500には、2台のサーバ（サーバ計算機）100a、100bと、コマンド送信用計算機300とが接続されている。またネットワーク500には、サーバ100aまたは100bからのサービスを受ける図示せぬクライアント（クライアント計算機）も接続されている。サーバ100a、100bには、それぞれ外部記憶装置としての（ハードディスク装置に代表される）ディスク装置200a、200bが接続されている。この計算機システムは共有記憶装置を持たない。

【0036】図1の計算機システムは、2台のサーバ100a、100bのうちの一方が処理を行い、処理を行っているサーバで障害が発生しても、もう一方のサーバが処理を引き継ぐことができるだけでなく、両サーバ100a、100bで共に障害が発生して、その後少なくとも一方が障害から復帰した場合にも、処理を引き継ぐサーバ、つまり最後に処理を行っていたサーバを正しく判定することが可能な仕組みを有していることに特徴がある。この仕組みについては後述する。

【0037】なお、本実施形態における「障害の発生」とは、ハードウェア障害、ソフトウェア障害などの他に、電源の切断などを含めて、処理の続行ができなくなる状態をいう。また、「障害からの復帰」とは、ハードウェア障害からの復帰、ソフトウェア障害からの復帰の他に、電源の投入などを含めて、処理の続行が行える状態になることをいう。

【0038】さて、処理の引き継ぎを可能とするためには、単に同じ処理を別のサーバで行うだけではなく、それまで処理を行っていたサーバから、処理の続行を行うのに必要な情報を受け取っておかなければならない。ところが本実施形態のように共有記憶装置を持たないシステムでは、当該共有記憶装置を介してのサーバ間の情報授受ができないことから、処理を行うサーバが、引き継ぎに必要な情報を他方のサーバに送らなければならない。しかし、この方式では、障害の発生時期と引き継ぎに必要な情報の転送時期との関係で、【従来の技術】の欄で述べたように、最後に処理を行っていたサーバを判定することが困難となる場合がある。そこで、最後に処理を行っていたサーバが容易に判定できる方法として、【従来の技術】で述べた「引き継ぎを1度に限定する方法」、或いは「時刻情報を使用する方法」が知られているが、自動運転が実現できないとか判定精度（完全性）が不十分であるといった問題がある。

【0039】ここで自動運転について詳述する。自動運

転とは、【従来の技術】の欄で述べたように、片方または両方のサーバに障害発生、或いは障害復帰がいつ起こっても、可能な限り処理を継続する運転方法を指す。自動運転を実現するためには、少なくとも次の動作（1）～（5）が行える必要がある。

【0040】（1）2台のサーバに障害が発生し、その後2台のサーバが復帰したとき、処理を行うことが可能か否かを判定し、可能な場合は処理を行うべきサーバを決定し、処理を開始する。もう1台のサーバは、引き続き可能な状態にする。「引き続き可能な状態にする」とは、処理を引き継ぐために必要な情報を、処理をしているサーバから、もう1台のサーバに全て送り、処理をしているサーバに障害が発生しても、もう1台のサーバが処理を続けられるようにすることを指す。

【0041】（2）2台のサーバに障害が発生し、その後1台のサーバが復帰したとき、復帰したサーバで処理を行うことが可能か否かを判定し、可能な場合は処理を開始する。

【0042】（3）1台のサーバが処理中で、もう1台のサーバは障害が発生している状態のとき、障害が発生しているサーバが復帰した際に、そのサーバを引き継ぎ可能な状態にする。

【0043】（4）1台のサーバが処理中で、もう1台のサーバが引き続き可能な状態にあるとき、処理中のサーバで障害が発生した場合は、引き継ぎを行う。

【0044】これに加え、次の動作（状態）は必ず考えなければならない。

（5）どのタイミングでも、障害が発生する。

【0045】これらの動作（状態）に基づき、図2に示す状態遷移図800が作成（用意）される。ここでは、サーバの状態を、マスタ（マスタ状態）M、シングルマスタ（シングルマスタ状態）SM、スレーブ（スレーブ状態）SL、及び停止（停止状態）Xの4つの状態に分類している。つまり本実施形態では、従来のマスタMを新たなマスタMとシングルマスタSMという2状態に分類し、マスタM、スレーブSL、及び停止Xの3つの状態にシングルマスタSMという状態が加えられた点に特徴がある。

【0046】サーバの状態は、まず、そのサーバが処理を実行しているか、実行していないかによって分けられ、
処理を行っている：
マスタM、シングルマスタSM
処理を行っていない：
スレーブSL、停止X
のようになる。

【0047】更に、「処理を行っている」「処理を行っていない」のそれぞれを、
処理を行っている：

行っている処理を引き継ぐ相手が存在する…マスタM

行っている処理を引き継ぐ相手が存在しない…シングル
マスタSM

処理を行っていない：

引き継ぎのための情報を受け取っている…スレーブSL
引き継ぎのための情報を受け取っていない…停止X
のように分類する。

【0048】明らかなように、サーバに障害が発生した場合、処理を行わず、引き継ぎもできないので、当該サーバの状態は停止Xの状態になる。

【0049】さて、図2の状態遷移図800では、一方のサーバ（例えばサーバ100a）の状態をA、もう一方のサーバ（例えばサーバ100b）の状態をBとすると、（A B）と表記している。（A B）と（B A）は異なる状態を表している。

【0050】図2の状態遷移図800で適用されるそれぞれの状態について説明する。まず（X X）は、両サーバが共に停止状態にある状態を示す。具体的には、どちらのサーバも処理も行っておらず、引き継ぎもできない状態を示す。

【0051】（SM SL）と（SL SM）とは、一方のサーバがSM（シングルマスタ）状態にあり、他方のサーバがSL（スレーブ）状態にある状態を示す。具体的には、引き継ぎ可能な状態にするために、SMの状態のサーバから、SLの状態のサーバに、処理を引き継ぐための情報を送っている状態を示す。この状態では、一方のサーバで処理は行われているが、他方のサーバでの引き継ぎはできない。

【0052】（SL M）と（M SL）とは、一方のサーバがM（マスタ）状態にあり、他方のサーバがSL（スレーブ）状態にある状態を示す。具体的には、一方のサーバで処理が行われ、且つ他方のサーバで引き継ぐことができる状態である。

【0053】（X SM）と（SM X）とは、一方のサーバがSM（シングルマスタ）状態にあり、他方のサーバがX（停止）状態にある状態を示す。具体的には、1台のサーバが処理を行っているが、もう1台のサーバに引き継ぎをするための情報を送ることができない状態である。つまり引き継ぎはできない。

【0054】次に、図2の状態遷移図800を参照して、上記動作（1）～（5）について説明する。
動作（1）：2台のサーバで共に障害が発生して停止している状態は、（X X）という状態である。ここで、両サーバが障害から復帰すると、図2の1-1-1→1-1-2、または1-2-1→1-2-2の状態遷移が行われる。

【0055】このように、両サーバが障害から復帰した場合には、まず、どちらかのサーバで処理を始める。そして、引き継ぎ可能な状態にするために、引き継ぎに必要な情報を、処理を行っているサーバから、もう1台のサーバに送る状態に遷移する（1-1-1または1-2

-1）。これが（SL SM）または（SM SL）状態である。

【0056】その後、引き継ぎが可能になったら、一方はスレーブ状態のまま、もう一方が（シングルマスタ状態から）マスタ状態という状態、つまり（SL M）または（M SL）の状態に移る（1-1-2または1-2-2）。この状態でも、引き継ぎに必要な情報はマスタからスレーブに送られる。

【0057】動作（2）：両サーバが障害発生（=停止）の状態（X X）で、一方のサーバだけが障害から復帰した場合は、2-1または2-2の遷移が行われる。これは、復帰したサーバを処理のできる状態（シングルマスタの状態）にするための状態遷移であり、（X SM）または（SM X）の状態に移る。

【0058】動作（3）：一方のサーバが処理中で、他方のサーバが障害発生（=停止）という状態は、（X SM）または（SM X）である。この（X SM）または（SM X）の状態、停止状態のサーバが障害から復帰した場合、それぞれ3-1-1→3-1-2、または3-2-1→3-2-2の状態遷移が行われる。

【0059】このように、（X SM）または（SM X）の状態、停止状態のサーバが障害から復帰した場合には、まず、一方はシングルマスタ状態のまま、もう一方は（停止状態から）スレーブ状態という状態に移り（3-1-1または3-2-1）、引き継ぎに必要な情報を、処理を行っているサーバ（シングルマスタ）から、もう一方のサーバ（スレーブ）に送る。これが（SL SM）または（SM SL）状態である。

【0060】その後、引き継ぎが可能になったら、動作（1）と同様に、一方はスレーブ状態のまま、もう一方が（シングルマスタ状態から）マスタ状態という状態、つまり（SL M）または（M SL）の状態に移る（3-1-2または3-2-2）。これは、それまで処理を行っていたサーバをマスタ、復帰したサーバをスレーブとして、マスタ、スレーブを確立するための状態遷移である。

【0061】動作（4）：引き継ぎが可能な状態ということは、（SL M）または（M SL）という状態にあることである。この状態で、スレーブ状態のサーバに障害が発生した場合は、（X SM）または（SM X）という状態に遷移する（4-1または4-2）。これは、処理はそのまま同じサーバが継続するものの、引き継ぎを行わない状態への状態遷移である。これに対し、マスタ状態のサーバに障害が発生した場合は、（SL M）または（M SL）の状態は（SM X）または（X SM）の状態に遷移する（4-3または4-4）。これは、それまでスレーブ状態だったサーバが処理を引き継ぐための状態遷移である。

【0062】動作（5）：サーバに障害が発生した場合、必ずそのサーバは停止状態（X）になってしまう。

これが5-1乃至5-8である。以下にそれぞれの動作を説明するまず、一方がマスタで、もう一方がスレーブの状態、つまり(SL M)または(M SL)の状態、両方のサーバに障害が発生した場合には、5-1または5-5の遷移をして(X X)の状態となる。

【0063】次に、シングルマスタ状態のサーバから引き継ぎを可能にするための情報をスレーブ状態のサーバに送っている(SL SM)または(SM SL)の状態、シングルマスタ状態のサーバに障害が発生した場合には、引き継ぎはできないため、5-2または5-6の遷移が行われて、どちらのサーバも停止状態という状態、つまり(X X)の状態となる。また、上記(SL SM)または(SMSL)の状態、両サーバに障害が発生した場合も、5-2または5-6の遷移が行われて、両サーバは停止状態になる。

【0064】また、上記(SL SM)または(SM SL)の状態、スレーブ状態のサーバに障害が発生した場合は、5-3または5-7の遷移が行われて(X SM)または(SM X)の状態に移る。この状態では、処理が残ったサーバ(シングルマスタ状態のサーバ)で続行される。

【0065】次に、一方のサーバがシングルマスタ状態で、もう一方のサーバが停止状態、つまり(X SM)または(SM X)の状態、シングルマスタ状態のサーバに障害が発生した場合には、5-4または5-8の遷移をして、両サーバ共に停止状態、つまり(X X)となる。

【0066】以上のように、図2に示した状態遷移図800には、自動運転を行うための動作(1)～(5)が全て含まれている。

【0067】さて、上記の動作(1)～(5)のうち、動作(3)～(5)については、状態の変わるサーバがどちらであるかが決まっているため、制御することは容易である。

【0068】しかし、動作(1)は、処理を引き継ぐサーバが自身であるか否かを判定する必要があり、動作(2)も、障害から復帰したサーバがそのまま処理を引き継いでもよいか否かを判定する必要がある。本実施形態は、この動作(1)及び(2)において処理を引き継ぐサーバを判定する手法を、先に述べたサーバ状態の分類を利用して工夫した点に特徴がある。この判定手法の詳細については別途説明する。ここでは、処理を引き継ぐサーバは、最後まで処理を行っていた方のサーバとなることから、そのことについて述べる。

【0069】まず、動作(1)において、処理を引き継ぐべきサーバである条件は、2台のサーバのうち、最後まで処理を行っていた方のサーバであることである。その理由を以下に説明する。

【0070】本実施形態において、それまでに処理を行っていないサーバが処理の続きを行うためには、処

理を行っていたサーバから、引き継ぎのための情報を受け取っておく必要がある。もし、最後に処理を行っていたサーバに障害が発生する直前に、当該サーバからもう一方のサーバに引き継ぎのための情報を送っていない場合は、もう一方のサーバでは処理の続きを行うことができない。そのため、最後に処理を行っていた方のサーバを、処理を引き継ぐべきサーバとする必要がある。

【0071】次に、最後に処理を行っていたサーバに障害が発生する前に、当該サーバからもう一方のサーバに引き継ぎのための情報を送っていた場合は、どちらのサーバでも処理の続きをすることが可能なように見える。しかし、障害発生直前に行っていた処理については、当該処理の引き継ぎに必要な情報を上記もう一方のサーバに送る前に、障害が発生している可能性がある。そのため、こちらの場合も、最後に処理を行っていた方のサーバを、処理を引き継ぐべきサーバとする必要がある。

【0072】以上から、動作(1)では、最後まで処理を行っていたサーバが処理を引き継ぐ必要がある。

【0073】次に動作(2)において、障害から復帰したサーバに処理をさせる条件は、動作(1)の場合と同じ理由により、当該障害復帰サーバが、2台のうち最後まで処理を行っていた方のサーバであることである。

【0074】さて、各サーバにおいて、自身が最後まで処理を行っていたサーバであるか否かを判定する手法として、[従来の技術]の欄で述べたような「時刻情報を使う方法」がある。しかし、「時刻情報を使う方法」は、時間というグローバルなものを暗黙に仮定し、各サーバの使う時計が常に同期しているという仮定のもとで成立しており、実際の時計は必ずしも常に同期しているとは限らないことから、完全性に問題がある。

【0075】このため本実施形態では、最後に処理を行っていたサーバをローカルな情報だけを使って判定する手法を適用している。但し、ローカルな情報には、次のような問題がある。それは、ローカルな情報は、当該情報を持つサーバが停止してしまうと、他のサーバから変更できないため、それぞれのサーバの持つ情報が矛盾してしまう可能性があるという問題である。

【0076】この問題について、処理を行っているサーバが、自分で処理を行っている旨の情報をローカルに持つ場合を例に述べる。この例では、処理をしているサーバは「自分が処理中」という情報を持っている。その後、このサーバに障害が発生し、相手サーバが処理を引き継いだ場合、もう一方のサーバも「自分が処理中」という情報を持つことになる。このとき、障害が発生したサーバの持つ「自分が処理中」という情報を書き換えることはできない。このような状態で、今度は、もう一方のサーバにも障害が発生し、その後、両方のサーバが障害から復帰したとする。このとき、両方のサーバが「自分が処理中」という情報を持つことになる。したがって、処理をしているサーバが「自分が処理中」という情

報をローカルに持つだけでは、どちらが最後に処理をしていたかを判定できない。

【0077】そこで本実施形態では、シングルマスタ状態を含む4つの状態に分類される各サーバの状態を、図2に示した状態遷移図800に基づいて制御し、後述するサーバ優先度と称する3つの状態をとり得る状態変数をローカルな情報として持つことにより、この問題を解決するようにしている。

【0078】ここで、再び図1を参照すると、サーバ100a、100bは同一構成を有している。即ちサーバ100a、100bは、いずれもクラスタ管理部110、処理・引き継ぎ部400、及びサーバ優先度処理部700を備えている。これら各部110、400、700は、サーバ100a、100bが所定のソフトウェアプログラムを読み込み実行することにより実現される機能手段である。ここでは、クラスタ管理部110を実現するためのソフトウェア（クラスタソフトウェア）と、処理・引き継ぎ部400を実現するためのソフトウェア（処理ソフトウェア）と、サーバ優先度処理部700を実現するためのサーバ優先度管理ソフトウェアとは、同一の記録媒体、例えばCD-ROMに記録して提供され、サーバ100a、100bの持つディスク装置200a、200bにインストールして用いられる。なお、上記各ソフトウェアがディスク装置200a、200bに予めインストールされていてもよく、また、それぞれ別々の記録媒体に記録されていてもよい。また、ネットワーク500を介してダウンロードされるものであっても構わない。

【0079】クラスタ管理部110は、図2に示した状態遷移図800に従う状態遷移制御を行う機能を有する。即ちクラスタ管理部110は、もう一方の（サーバ上で動作する）クラスタ管理部110とネットワーク500を通じて通信することにより、予め定められた状態遷移図800に基づき、2台のサーバ100a、100bの状態遷移制御を行う。またクラスタ管理部110は、状態遷移の際には、処理・引き継ぎ部400とサーバ優先度処理部700とに、最新の遷移状態を示す状態変化情報901を送る。

【0080】またクラスタ管理部110は、サーバ100a及び100bの両サーバが停止状態(X)にある状態で、自身が動作するサーバ（自サーバ）が障害から復帰した際には、サーバ優先度処理部700に自サーバの障害復帰を示す状態変化情報901を送る。更にクラスタ管理部110は、相手サーバのクラスタ管理部110との通信により、相手サーバの障害発生と、障害からの復帰とを検出し、状態遷移を行う。なおクラスタ管理部110は、自サーバの障害復帰時には、復帰後所定時間を待って、相手サーバも障害から復帰したか否かを当該相手サーバとの通信により調べ、しかる後に状態変化情報901を送るようにしている。

【0081】処理・引き継ぎ部400は、クラスタ管理部110から得られる（最新の遷移状態を示す）状態変化情報901に応じて、処理と処理引き継ぎのための動作の開始/停止の制御を行う。処理・引き継ぎ部400は、処理制御部410と、引き継ぎ制御部420とから構成される。

【0082】処理制御部410は処理の開始/停止を制御し、引き継ぎ制御部420は、処理の引き継ぎのための動作の開始/停止を制御する。処理とは、例えばDBMS（データベースマネジメントシステム）等のアプリケーションプログラムの実行である。

【0083】サーバ優先度処理部700は、自サーバ100i（iはaまたはb）のディスク装置200i（に確保された優先度記録領域）に自サーバ100iの優先度（サーバ優先度）210を記録し、当該優先度210を相手サーバ100j（jはaまたはb、但しi≠j）のディスク装置200jに記録された優先度210と比較する。このサーバ優先度210は最後まで処理を行っていたサーバを判定するために必要な情報であり、1、2、3の3状態のいずれかをとり。

【0084】サーバ優先度処理部700は、状態書き込み部710と、比較処理部720とから構成される。状態書き込み部710は、クラスタ管理部110から得られる状態変化情報901に応じて、自サーバ100iのディスク装置200i（に確保された優先度記録領域）に記録されている自サーバ100iの優先度210を更新する。

【0085】比較処理部720は、クラスタ管理部110から自サーバの障害復帰を示す状態変化情報901を受け取った場合、相手サーバの比較処理部720との間でネットワーク500を介して優先度210の授受を行う。そして比較処理部720は、自サーバ100iの優先度210と相手サーバ100jの優先度210とを比較することで、自サーバ100iの方が優先度が高いか否かの判定を行う。この判定の詳細は後述する。比較処理部720での優先度比較・判定結果は、クラスタ管理部110で行われる、自サーバ100iが最後に処理をしたサーバであるか否か（つまり自サーバ100iがマスタとなるサーバであるか否か）の判定に用いられる。

【0086】コマンド送信用計算機300は、ユーザから要求された強制開始命令を、ネットワーク500を介して、サーバ100a、100bのうち当該命令の指定するサーバのクラスタ管理部110に送信する。この強制開始命令は、指定のサーバで強制的に処理を開始させる命令である。

【0087】ここで、強制開始命令を用いた処理の強制開始機能を導入する技術的背景について述べる。サーバ100a及び100bの両サーバに障害が発生し、その後、一方のサーバ100iだけが障害から復帰した場合、つまり上記動作(2)の場合、その障害復帰サーバ

100iは処理を行うべきサーバではないと判定されることがあり得る。この場合の動作としては、もう一方の、本来処理を行うべきサーバが復帰するまで待つことが考えられる。

【0088】しかし、もう一方のサーバが復帰するまで処理が停止したままであるよりは、処理の続きができなくても構わないので、先に障害から復帰したサーバ100iで処理を再開した方が都合がよい場合も当然あり得る。

【0089】そこで本実施形態では、このような場合に、先に障害復帰したサーバで強制的に処理を開始できる強制開始モードと、本来処理を行うべきサーバが障害から復帰するまで待つ待機モードの2種のモードの中から、いずれか一方のモードが選択指定できるようにしている。ここでは、通常は待機モードに自動設定され、コマンド送信用計算機300から強制開始命令が与えられた場合だけ、強制開始モードに切り替えられる構成を適用している。

【0090】次に、サーバ100a、100bの動作の詳細について、サーバ優先度処理部700の動作を中心に説明する。処理・引き継ぎ部400は、クラスタ管理部110から最新の遷移状態を示す状態変化情報901が送られると、当該状態変化情報901を受け取り、当該状態変化情報901の示す最新の状態に従って、以下に述べる、処理の開始もしくは停止、または引き継ぎのための動作の開始もしくは停止を行う。

【0091】まず、処理の開始、停止については、前述したマスタ、シングルマスタ、スレーブ、停止の各状態の説明に従い、処理・引き継ぎ部400内の処理制御部410が次のように動作する。

【0092】自サーバ100iがマスタ(M)、またはシングルマスタ(SM)の状態となったときは処理を行う。これに対し、自サーバ100iがスレーブ(SL)、または停止状態(X)となったときは処理を行わない。

【0093】次に、引き継ぎのための動作の開始、停止については、処理・引き継ぎ部400内の引き継ぎ制御部420が次のように動作する。一方がシングルマスタまたはマスタ、もう一方がスレーブの状態になった状態では、処理引き継ぎのための情報の授受をネットワーク500を介して行う。ここでは、シングルマスタまたはマスタ側の引き継ぎ制御部420が引き継ぎのための情報の送り手となり、スレーブ側の引き継ぎ制御部420が当該情報の受け手となる。他の状態の組み合わせでは、引き継ぎのための情報授受は行わない。

【0094】次に、サーバ優先度処理部700の動作の詳細について、サーバ優先度処理部700を構成する状態書き込み部710と比較処理部720のそれぞれの動作に分けて順に説明する。まず、サーバ優先度処理部700内の状態書き込み部710は、クラスタ管理部11

0から最新の遷移状態を示す状態変化情報901が送られた場合、当該状態変化情報901の示す最新の状態に従って、自サーバ100iのディスク装置200i(に確保された優先度記録領域)に記録されているサーバ優先度210を変更するサーバ優先度変更(記録)処理を、図3のフローチャートに従って次のように行う。

【0095】自サーバ100iの状態がシングルマスタ状態(SM)に遷移したときは、1に変更する(ステップS1、S2)。自サーバ100iの状態がマスタ状態(M)に遷移したときは、2に変更する(ステップS1、S3)。自サーバ100iの状態がスレーブ状態(SL)に遷移したときは、3に変更する(ステップS1、S4)。自サーバ100iの状態が停止状態(X)に遷移したときは、変更しない(ステップS1、S5)。

【0096】ここで、サーバ優先度210の初期値は、サーバ100aと100bとで異なり、一方のサーバでは2に、もう一方のサーバでは3に設定される。このサーバ優先度210の初期設定は、自動運転を始める前の処理で行われる。

【0097】次に、サーバ優先度処理部700内の比較処理部720の動作、つまり最後に処理を行っていたサーバをクラスタ管理部110にて判定するのに必要な優先度比較・判定動作について、強制開始を行わない場合と行う場合とに分けて順次説明する。

【0098】まず、強制開始を行わない場合の動作を、図4のフローチャートを参照して説明する。図2に示した状態遷移図800において、最後に処理を行っていたサーバを判定する必要があるのは、1-1-1→1-1-2または1-2-1→1-2-2の遷移の場合と、2-1または2-2の遷移の場合の2つ、つまり2台のサーバ100a、100bに障害が発生して共に停止状態(X X)となっているときに、両サーバが障害から復帰した場合と、いずれか一方のサーバが障害から復帰した場合である。なお、以下の説明では、1-1-1→1-1-2、1-2-1→1-2-2を、その共通部分をとって1-1、1-2で表現する。サーバ100a、100bのクラスタ管理部110は、(X X)の状態で自サーバが障害復帰すると、その旨を示す状態変化情報901を自サーバのサーバ優先度処理部700に送る。

【0099】サーバ優先度処理部700内の比較処理部720は、自サーバのクラスタ管理部110から自サーバが障害復帰したことを示す状態変化情報901が送られた場合、上記1-1もしくは1-2の遷移、または2-1もしくは2-2の遷移のために、自サーバが最後に処理を行っていたか否かの判定(自サーバが処理を引き継ぐか否かの判定)に必要なサーバ優先度210の比較・判定が要求されたものと判断し、当該比較・判定を次のように行う。

【0100】1-1もしくは1-2の遷移となる場合：

(1) 比較処理部 720 は、相手サーバ (サーバ 100j とする) の比較処理部 720 とネットワーク 500 を介して通信を行ってサーバ優先度 210 を授受し、自サーバ (サーバ 100i とする) と相手サーバ 100j とで、どちらのサーバ優先度 210 が高いかを比較・判定する (ステップ S11, S12)。ここでサーバ優先度 210 は 1 が最も高く、2, 3 の順に低くなる。つまり、サーバ優先度 210 は 1 で最高優先度を、2 で 2 番目の優先度を、3 で最低優先度を示す。なお、自サーバ 100i のサーバ優先度 210 が 1 の場合には、相手サーバ 100j のサーバ優先度 210 を取得することなく、自サーバ 100i の方が優先度が高いと判定することも可能である。この理由については後述する。

【0101】(2) 比較処理部 720 は、サーバ優先度 210 の比較・判定結果、つまり自サーバ 100i の方が優先度が高いか否かのサーバ優先度判定結果を、自サーバのクラスタ管理部 110 から (X X) 状態で送られた状態変化情報 901 に対する応答として、当該クラスタ管理部 110 に通知する (ステップ S13)。

【0102】(3) クラスタ管理部 110 は、比較処理部 720 からサーバ優先度判定結果を受け取ると、当該判定結果と相手サーバ 100j も障害復帰しているか否かと状態遷移図 800 とに基づき、図 5 のフローチャートに従って次のように状態遷移する。

【0103】まずクラスタ管理部 110 は、比較処理部 720 からのサーバ優先度判定結果により自サーバ 100i の方がサーバ優先度 210 が高いと判定できる場合 (ステップ S21, S22)、自サーバ 100i が最後に処理を行っていたサーバであると判断する。しかも、相手サーバ 100j も障害復帰している場合 (ステップ S23)、クラスタ管理部 110 は、最終的に自サーバ 100i がマスタとなるために、まずシングルマスタ状態 (SM) となり、相手サーバ 100j がスレーブ状態 (SL) となるように遷移を行う (ステップ S24)。これに対し、相手サーバ 100j の方がサーバ優先度 210 が高い場合には (ステップ S25)、自サーバ 100i は最後に処理を行っていたサーバではなく、最終的に相手サーバ 100j がマスタとなるものとして、自サーバ 100i がスレーブ (SL) 状態となり、相手マスタ 100j がシングルマスタ状態 (SM) となるように遷移を行う (ステップ S26)。つまり、図 2 の 1-1-1 (SL SM) または 1-2-1 (SM SL) の状態遷移が行われる。なお、相手サーバ 100j が障害復帰していない場合には (ステップ S23)、後述する 2-1 (X SM) もしくは 2-2 (SM X) の遷移となる (ステップ S27)。

【0104】(4) クラスタ管理部 110 は、この状態遷移結果、つまり最新の遷移状態 (SL SM) または (SM SL) を示す状態変化情報 901 を処理・引き継ぎ部 400 及びサーバ優先度処理部 700 に送る。こ

の際の処理・引き継ぎ部 400 (内の処理制御部 410 及び引き継ぎ制御部 420) と、サーバ優先度処理部 700 内の状態書き込み部 710 の動作内容は、先の説明から明らかであり、(サーバ 100i または 100j に再び障害が発生しないならば) 最終的にはサーバ優先度 210 の高い方のサーバがマスタとなるように遷移する。つまり、図 2 の 1-1-2 (SL M) または 1-2-2 (M SL) の状態に遷移する。

【0105】2-1 もしくは 2-2 の遷移となる場合：

(1) 比較処理部 720 は、相手サーバ 100j の比較処理部 720 とネットワーク 500 を介して通信を行ってサーバ優先度 210 を授受し、自サーバ 100i と相手サーバ 100j とで、どちらのサーバ優先度 210 が高いかを比較する。ところが、2-1 もしくは 2-2 の遷移となる場合には、相手サーバ 100j はまだ停止状態 (X) にあることから、相手サーバ 100j のサーバ優先度 210 を取得することができない。しかし、自サーバ 100i のサーバ優先度 210 が 1 であれば、後述するように自サーバ 100i の方が優先度が高いと判断できる。

【0106】(2) そこで比較処理部 720 は、相手サーバ 100j が停止状態にあるために当該相手サーバ 100j のサーバ優先度 210 が取得できない場合には (ステップ S11)、自サーバ 100i のサーバ優先度 210 が 1 (つまり最高優先度) であるか否かを判定する (ステップ S14)。もし、自サーバ 100i のサーバ優先度 210 が 1 の場合には、比較処理部 720 は相手サーバ 100j のサーバ優先度 210 との比較を行うことなく、自サーバ 100i の方がサーバ優先度 210 が高いことを示すサーバ優先度判定結果を、自サーバ 100i のクラスタ管理部 110 に通知する (ステップ S15, S13)。

【0107】これに対し、自サーバ 100i のサーバ優先度 210 が 1 でない場合には (ステップ S14)、自サーバ 100i のサーバ優先度 210 だけでは、自サーバの方が優先度が高いか否かを判定できないとして、判定不可 (不定) を示す優先度判定結果を自サーバ 100i のクラスタ管理部 110 に通知する (ステップ S16, S13)。なお、自サーバ 100i のサーバ優先度 210 が 3 の場合には、自サーバ 100i の方が優先度が低い、つまり相手サーバ 100j の方が優先度が高いと判定できる。しかし、この例のように相手サーバ 100j が停止状態にある場合には状態遷移できない。そこで本実施形態では、相手サーバ 100j が障害復帰していない場合、自サーバ 100i のサーバ優先度 210 が 1 でない限り、全て判定不可として扱うようにしている。

【0108】(3) クラスタ管理部 110 は、比較処理部 720 からサーバ優先度判定結果を受け取ると、当該判定結果と相手サーバ 100j も障害復帰しているか否

かと状態遷移図800とに基づき、図5のフローチャートに従って次のように状態遷移する。

【0109】まずクラスタ管理部110は、相手サーバ100jが障害復帰していない状態で、自サーバ100iの方がサーバ優先度210が高い場合には（ステップS21～S23）、自サーバ100iが最後に処理を行っていたサーバであるとして、状態遷移図800に従い、自サーバ100iがシングルマスタになるように遷移を行う（ステップS27）。これに対し、判定結果が判定不可を示す場合（相手サーバ100jが障害復帰しておらず、且つ自サーバ100iのサーバ優先度210が1でない場合）には（ステップS25）、相手サーバ100jが障害復帰するまで状態遷移を行わない。

【0110】図2に示した状態遷移図800に従い、上記の判定に基づいてサーバ状態が遷移したとすると、サーバ優先度210は図6のように遷移する。この図では、一方のサーバの状態をA、サーバ優先度をa、もう一方のサーバの状態をB、サーバ優先度をbとしたとき、(A B)と表記し、その下に[a b]と表記している。(A B)と(B A)は異なる状態を表す。[a b]と[b a]は異なる場合を表す。

【0111】以下で、上述の判定法によって、最後に処理を行っていたサーバを正しく判定できる理由について説明する。

命題1 サーバ優先度210の高いサーバが最後に処理を行っていたサーバである。

命題2 あるサーバのサーバ優先度210が1であれば、そのサーバは必ず最後に処理を行っていたサーバである。

【0112】この2つの命題1、2を証明する。そのためには、次の3つの補題1～3が成り立てばよい。

【0113】補題1

仮定：

・ある時点で、自サーバ100iのサーバ優先度210に1がセットされている。その後、自サーバ100iのサーバ優先度210は変更されておらず、相手サーバ100jがマスタ状態またはシングルマスタ状態になっていない。

・相手サーバ100jのサーバ優先度210が1ではない。

結論：自サーバ100iのサーバ優先度210が1であるとき、最後に処理を行ったのは自サーバ100iである。

【0114】補題2

仮定：

・ある時点で、自サーバ100iのサーバ優先度210に2がセットされている。その後、自サーバ100iのサーバ優先度210は変更されておらず、相手サーバ100jがマスタ状態またはシングルマスタ状態になっていない。

・相手サーバ100jのサーバ優先度210が3である。

結論：自サーバ100iのサーバ優先度210は2であり、相手サーバ100jのサーバ優先度210が3であるとき、最後に処理を行ったのは自サーバ100iである。

【0115】補題3

仮定：両サーバ100i、100jが停止した状態である。

結論：サーバ優先度210は、互いに異なる値をとる。

【0116】命題1の証明：まず、補題1、2、3の結論を使い、命題1を証明する。命題1を証明するには、サーバ優先度210の全ての組み合わせにおいて、優先度の高いサーバが最後に処理を行っていたことを証明できればよい。

【0117】補題3から、両サーバ100i、100jが停止した状態のとき、サーバ優先度210は互いに異なっているから、(1, 2, 3の3状態を持つ)サーバ優先度210の組み合わせは、(1, 2)(1, 3)(2, 3)の3通りである。

【0118】補題1から、(1, 2)(1, 3)についてはサーバ優先度210が1のサーバが最後に処理を行っている。次に補題2から、(2, 3)については、サーバ優先度210が2のサーバが最後に処理を行っている。よって、全ての場合において、サーバ優先度210の高いサーバが最後に処理を行っているといえる。

【0119】命題2の証明：次に、補題1、3の結論を使って命題2を証明する。命題2の仮定より、自サーバ100iのサーバ優先度210は1である。このとき、補題3より、相手サーバのサーバ優先度は1ではない。このとき補題1より、自サーバ100iは必ず最後に処理を行っていたサーバである。このように、補題1、2、3が成り立てば、命題1及び2は成り立つ。

【0120】次に、補題1、2、3それぞれについて、仮定が成り立てば結論が成り立つことの証明を行い、最後に、補題の仮定が成り立つことを説明する。

【0121】補題1の仮定から、結論が成り立つことの証明：自サーバ100iのサーバ優先度210に1がセットされたということは、その時点で、自サーバ100iはシングルマスタ状態であったことを意味する。その後、相手サーバ100jは1度もマスタあるいはシングルマスタになっていないということは、相手サーバ100jはその時点から1度も処理を行っていないことである。よって、最後に処理を行ったのは、自サーバ100iである。

【0122】自サーバ100iのサーバ優先度210に1がセットされた時点より後に、当該優先度210が変更されていないということは、自サーバの現在の優先度210は1である。仮定から、現在の相手サーバ100jのサーバ優先度210は1ではない。よって、自サー

サーバ100iのサーバ優先度210が1であり、相手サーバ100jのサーバ優先度210が1ではないとき、最後に処理を行ったのは自サーバ100iである。

【0123】補題2の仮定から、結論が成り立つことの証明：自サーバ100iのサーバ優先度210に2がセットされたということは、その時点で、自サーバ100iはマスタ状態であったことを意味する。マスタ状態になるためには、相手サーバ100jは必ずスレーブ状態である必要があり、したがって、その時点での相手サーバ100jのサーバ優先度210は3である。

【0124】その後、相手サーバ100iは1度もマスタあるいはシングルマスタになっていないということは、相手サーバ100jはその時点から1度も処理を行っていないということである。よって、最後に処理を行ったのは、自サーバ100iである。

【0125】また、相手サーバ100jのサーバ優先度210は、マスタ或いはシングルマスタにならなければ変わらないので、相手サーバ100jのサーバ優先度210は3のままである。

【0126】また、自サーバ100iのサーバ優先度210に2がセットされた時点より後に当該優先度210が変更されていないということは、自サーバ100iの現在の優先度210は2である。

【0127】よって、自サーバ100iのサーバ優先度210が2であり、相手サーバ100jのサーバ優先度210が3であるとき、最後に処理を行ったのは自サーバ100iである。

【0128】補題3の仮定から、結論が成り立つことの証明：両サーバ100i、100jが停止した状態のとき、サーバ優先度210は常に互いに異なる状態をとることを証明するには、サーバ優先度210が常に互いに異なる状態をとることを証明すればよい。そのためには、次の4つの仮定3-1～3-4を証明すればよい。

【0129】仮定3-1：サーバ優先度210の初期値は互いに異なっている。

仮定3-2

片方のサーバのサーバ優先度210が1のとき、もう片方のサーバのサーバ優先度210は1にならない。

仮定3-3

片方のサーバのサーバ優先度210が2のとき、もう片方のサーバのサーバ優先度210は2にならない。

仮定3-4

片方のサーバのサーバ優先度210が3のとき、もう片方のサーバのサーバ優先度210は3にならない。これらが証明できれば、片方のサーバのサーバ優先度210が何であっても、もう片方のサーバのサーバ優先度210は同じにならないことがいえる。

【0130】仮定3-1は、サーバ優先度210の動作より明らかである。以下で、仮定3-2～3-4を証明する。

仮定3-2の証明：片方のサーバ（サーバ100pとする）のサーバ優先度210が1になったとき、もう片方のサーバ（サーバ100qとする）は停止状態である。この状態で、サーバ100pのサーバ優先度210を変えずにサーバ100qのサーバ優先度210を1にするには、サーバ100qをシングルマスタ状態にしなければならない。そのためには、両サーバが一旦、停止状態に遷移し、更にサーバ100pがスレーブ状態、サーバ100qが停止状態という状態に遷移しなければならない。しかし、この時点で、サーバ100pのサーバ優先度210は1であり、これは両サーバが停止した状態から、マスタスレーブ状態への状態遷移の決定方法に矛盾する。よって、自サーバのサーバ優先度210を変更せずに相手サーバがシングルマスタ状態になることは不可能である。つまり、片方のサーバ100pのサーバ優先度210が1のとき、もう片方のサーバ100qのサーバ優先度210は1にならない。

【0131】仮定3-3の証明：片方のサーバ100pのサーバ優先度210が2になったとき、もう片方のサーバ100qは必ず3になる。よって、片方のサーバ100pのサーバ優先度が2のとき、もう片方のサーバ100qのサーバ優先度は2にならない。

【0132】仮定3-4の証明：片方のサーバ100pのサーバ優先度210が3になったとき、もう片方のサーバ100qはマスタ状態である。この状態で、サーバ100pのサーバ優先度210を変えずにサーバ100qのサーバ優先度210を3にするには、サーバ100qをスレーブ状態にしなければならない。そのためには、両サーバ100p、100qが一旦停止状態に遷移し、さらにサーバ100pが停止状態、サーバ100qがスレーブ状態という状態に遷移しなければならない。しかし、このときの両者のサーバ優先度210は3と2であり、サーバ100qの方が優先度210が高い。これは両サーバ100i、100jが停止した状態から、マスタスレーブ状態への状態遷移の決定方法に矛盾する。よって、自サーバのサーバ優先度210を変更せずに相手サーバがスレーブ状態になることは不可能である。つまり、片方のサーバ100pのサーバ優先度210が3のとき、もう片方のサーバ100qのサーバ優先度210は3にならない。このように補題1、2、3について、仮定から、結論が成り立つことが証明できた。

【0133】次に補題1、2の仮定が成り立つことを説明する。

補題1の仮定が成り立つことの証明：自サーバ100iのサーバ優先度210に1がセットされたとき、相手サーバ100jのサーバ優先度210が1でなければ、自サーバ100iのサーバ優先度210を変えずに、相手サーバ100jがマスタ状態またはシングルマスタ状態になることができないことを、図2の状態遷移図800を参照して説明する。

【0134】自サーバ100iのサーバ優先度210が1にセットされた後に、相手サーバ100jがマスタ状態になるためには、自サーバ100iはスレーブ状態になる必要があり、そのとき自サーバ100iのサーバ優先度210は3に変更されてしまう。つまり、自サーバ100iのサーバ優先度210を変更せずに相手サーバ100jがマスタ状態になることは不可能である。よって、自サーバ100iのサーバ優先度210が1にセットされた後、相手サーバ100jはマスタ状態になっていない。

【0135】自サーバ100iのサーバ優先度210が1にセットされたより後に、相手サーバ100jがシングルマスタになるためには、図2の状態遷移図800から、一旦、相手サーバ100jがマスタ、自サーバ100iがスレーブ、もしくはその逆に、相手サーバ100jがスレーブ、自サーバ100iがマスタという状態になるか、或いは、一旦両サーバ100i, 100j共に停止状態になって、その後に相手サーバ100jがシングルマスタにならなければならない。

【0136】ところが、一方がマスタで、他方がスレーブになる場合は、自サーバ100iのサーバ優先度210が変わってしまうので不可能である。また、両サーバ共に停止状態になり、相手サーバ100jがシングルマスタになるためには、判定方法の定義から、相手サーバ100jのサーバ優先度210が1でなければならない。これは仮定に反する。よって、自サーバ100iのサーバ優先度210に1がセットされた後、相手サーバ100jはシングルマスタ状態になっていないといえる。以上のことから、自サーバ100iのサーバ優先度210を変えずに、相手サーバ100jがマスタ状態またはシングルマスタ状態になることはできない。

【0137】補題2の仮定が成り立つことの証明：自サーバ100iのサーバ優先度210が2にセットされたとき、自サーバ100iのサーバ優先度210を変えずに、相手サーバ100jがマスタ状態またはシングルマスタ状態になることができないことを、図2の状態遷移図800を参照して説明する。

【0138】自サーバ100iのサーバ優先度210が2にセットされた時点での、自サーバ100iの状態は必ずマスタ状態であり、相手サーバ100jの状態は必ずスレーブ状態である。

【0139】自サーバ100iのサーバ優先度210が2にセットされた後に、相手サーバ100jがマスタ状態になるためには、自サーバ100iはスレーブ状態になる必要があり、そのとき自サーバ100iのサーバ優先度210は3に変更されてしまう。つまり、自サーバ100iのサーバ優先度210を変更せずに相手サーバ100jがマスタ状態になることは不可能である。よって、自サーバ100iのサーバ優先度210が2にセットされた後、相手サーバ100jはマスタ状態になって

いない。

【0140】自サーバ100iのサーバ優先度210に2がセットされた後に、自サーバ100iのサーバ優先度210を変えずに相手サーバ100jがシングルマスタ状態になることは可能である。しかし、その後、自サーバ100iのサーバ優先度210を変えずに、相手サーバ100jのサーバ優先度210を3にするためには、図2の状態遷移図800から、両サーバ100i, 100jが一旦停止状態に遷移し、更に自サーバ100iが停止状態、相手サーバ100jがスレーブ状態という状態に遷移しなければならない。ところが、この時点での両者のサーバ優先度210は自サーバ100iが2で相手サーバ100jが1であり、これは両サーバ100i, 100jが停止した状態から、マスタスレーブ状態への状態遷移の決定方法に矛盾する。よって、自サーバ100iのサーバ優先度210を変えずに、相手サーバ100jがシングルマスタ状態になることは不可能である。つまり、自サーバ100iのサーバ優先度210が2にセットされた後、相手サーバ100jはシングルマスタ状態になっていない。

【0141】以上により、自サーバ100iのサーバ優先度210が2にセットされたより後に、相手サーバ100jはマスタまたはシングルマスタになっていない。よって、両サーバ100i, 100jが停止した状態のとき、どちらかのサーバのサーバ優先度210が2で、もう一方のサーバ優先度210が3であった場合、サーバ優先度210が2のサーバ、つまり優先度が高い方のサーバが最後に処理をしていたサーバであるといえる。

【0142】次に、強制開始を行う場合の動作を説明する。図7に、強制開始がある場合の状態遷移及びサーバ優先度210の遷移を示す。図7に示す強制開始がある場合の遷移が、図6に示した強制開始がない場合と相違する点は、(X X)の状態からの2-1または2-2の遷移(図7において太線の矢印で示す遷移)で、サーバ優先度210の低い方のサーバを、強制的にシングルマスタにする遷移があるということである。

【0143】そのため、(X X), (X SM), (SM X)の3状態のとき、[11]というサーバ優先度210を持つ場合がある(図において※印が付されている優先度)。

【0144】例えば、(X X)の状態で、サーバ優先度が[1 3]や、[1 2]のとき、(X SM)へ強制開始によって状態遷移を行うことができる。そのとき、サーバ優先度は[1 1]になる。この状態でシングルマスタ状態のサーバに障害が発生した場合、(X X)で[1 1]になる。

【0145】このように、強制開始の結果、両サーバのサーバ優先度210が等しい状態が発生した場合、優先度の高いサーバを判定することができなくなる。したがって、両サーバが障害から復帰した場合に、最後に処理

を行っていたサーバを判定することができなくなる。この場合は、もう 1 度強制開始を行う必要がある。

【0146】また、強制開始の結果、上記の如く(X X)のときに[1 1]という状態があるため、一方のサーバが障害から復帰しても、2-1または2-2の遷移を自動的に行うことができなくなる。2-1または2-2の遷移をさせるためには、必ずもう 1 度強制開始を行う必要がある。

【0147】以上に述べたように、本実施形態においては、サーバの状態を、処理を行い、且つ行っている処理を引き継ぐ相手が存在する「マスタ」と、処理は行うが、行っている処理を引き継ぐ相手が存在しない「シングルマスタ」と、処理を行っていないが、引き継ぎのための情報を受け取っている「スレーブ」と、処理を行っておらず、且つ引き継ぎのための情報を受け取っていない「停止」の 4 つの状態に分類して作成された状態遷移図 800 に従って、サーバ 100a、100b の状態遷移を行うと共に、この状態の遷移で決まるサーバ優先度 210 という状態変数をローカルなディスク装置 200a、200b に記録し、当該状態遷移と状態変数(サーバ優先度 210)を用いて、最後まで処理を行っていたサーバを判定するようにしたので、次のような効果を得ることができる。まず、強制開始を使わない場合、サーバ 100a、100b の片方または両方(つまり少なくとも一方のサーバ)に、どのようなタイミングで障害発生、或いは障害復帰が起こっても、サーバ 100a、100b に障害が発生し、復帰したときに、当該サーバ自身が処理を行うべきサーバであるか否かを判定することができる。

【0148】特に本実施形態においては、前記した自動運転を行うための動作(1)～(5)を、2 台のサーバ 100a、100b 内で状態遷移のできるクラスタ管理部 110 を使うことで実現し、動作(1)(2)で必要なサーバの判定を、状態遷移の度にサーバ優先度処理部 700 内の状態書き込み部 710 がローカルなディスク装置に遷移状態に対応したサーバ優先度 210 を記録して、両サーバ 100a、100b のサーバ優先度 210 をサーバ優先度処理部 700 内の比較処理部 720 で比較することによって実現しているため、共有ディスク装置や時刻情報を使うことなく、自動運転を実現できる。

【0149】また、強制開始を使う場合、動作(2)において、復帰したサーバが処理を引き継ぐサーバであると判定できない場合でも、必要に応じて強制的に処理を引き継ぐことができる。この場合においても、両サーバ

が(X X)で[1 1]という特別な状態に遷移した場合を除き、自動的に処理を継続することができる。また、上記特別な状態の場合も、再度強制開始を用いて処理を継続することができる。

【0150】

【発明の効果】以上詳述したように本発明によれば、所定の状態遷移図に従う状態の遷移で決まるサーバ優先度という状態変数をローカルな記憶装置に記録し、当該状態遷移と状態変数を用いて、最後まで処理を行っていたサーバを判定するようにしたので、処理引き継ぎの回数の制限をなくして自動運転を実現すると共に、判定精度を向上することができる。

【図面の簡単な説明】

【図 1】本発明の一実施形態に係る高可用性計算機システムの構成を示すブロック図。

【図 2】同実施形態で適用される状態遷移図 800 を示す図。

【図 3】図 1 中の状態書き込み部 710 によるサーバ優先順位変更処理を説明するためのフローチャート。

【図 4】図 1 中の比較処理部 720 によるサーバ優先度判定処理を説明するためのフローチャート。

【図 5】図 1 中のクラスタ管理部 110 による障害復帰時状態遷移処理を説明するためのフローチャート。

【図 6】各サーバがローカルに持つサーバ優先度を、強制開始がない場合について、図 2 の状態遷移図 800 と対応付けて示す図。

【図 7】各サーバがローカルに持つサーバ優先度を、強制開始がある場合について、図 2 の状態遷移図 800 と対応付けて示す図。

【符号の説明】

100a、100b…サーバ

110…クラスタ管理部(状態遷移手段、最終処理計算機判定手段)

200a、200b…ディスク装置(記憶装置)

210…サーバ優先度

300…コマンド送信用計算機

400…処理・引き継ぎ部

410…処理制御部

420…引き継ぎ制御部

500…ネットワーク

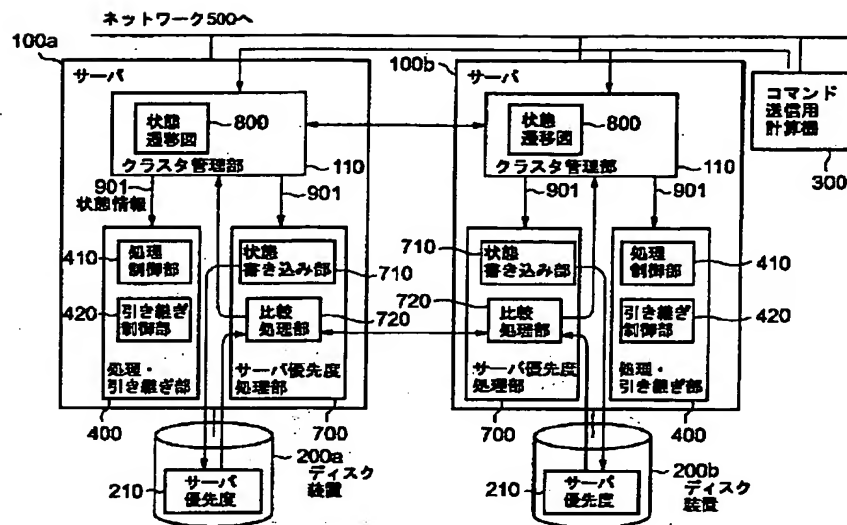
700…サーバ優先度処理部

710…状態書き込み部

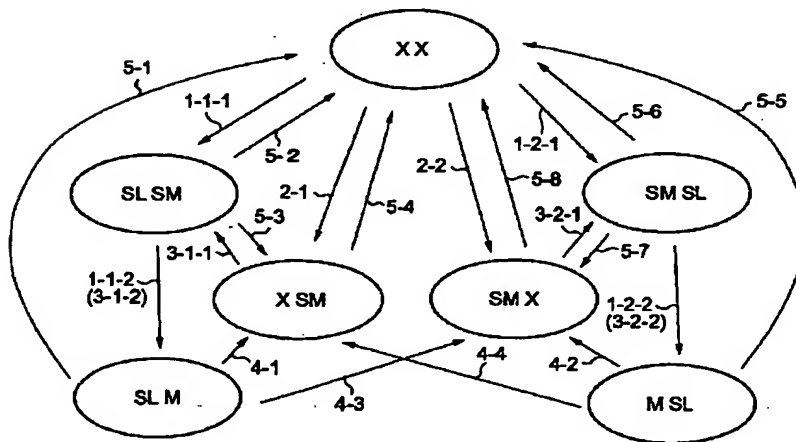
720…比較処理部(優先度判定手段)

901…状態変化情報

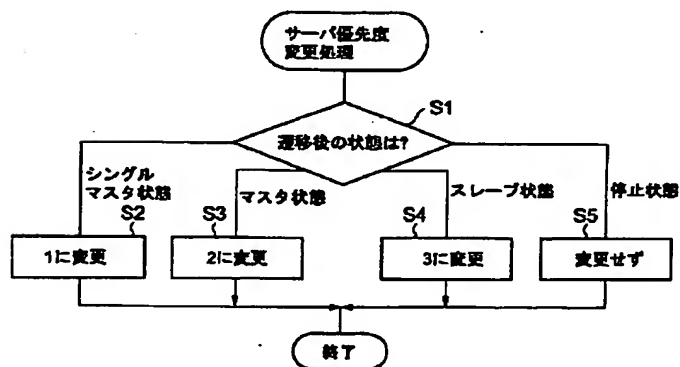
【図1】



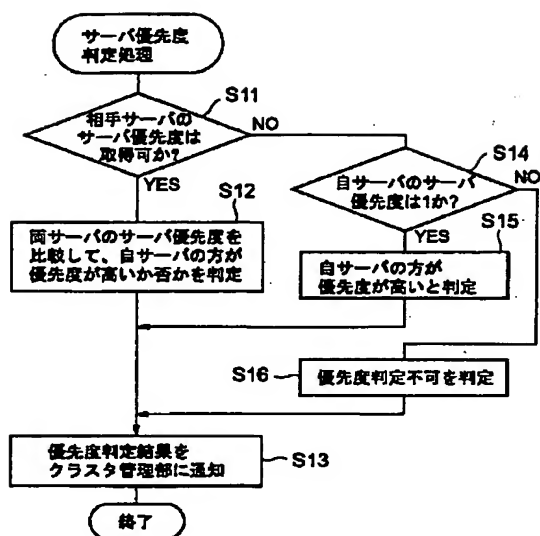
【図2】



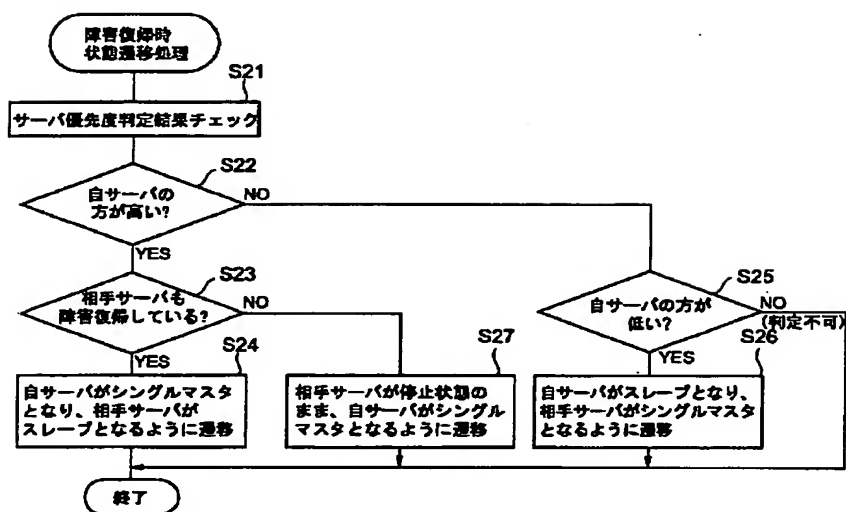
【図3】



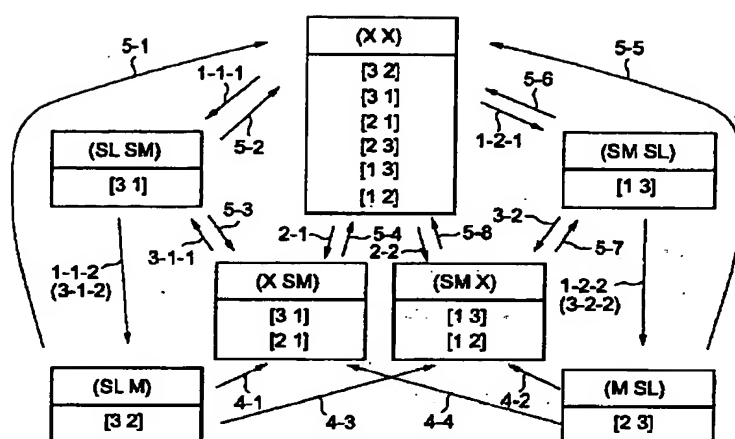
【図4】



【図5】



【図 6】



【図 7】

